ORIGINAL ARTICLE

Parentage and sibship exclusions: higher statistical power with more family members

J Wang

Institute of Zoology, Zoological Society of London, London, UK

Parentage exclusion probabilities are now routinely calculated in genetic marker-assisted parentage analyses to indicate the statistical power of the analyses achievable for a given set of markers, and to measure the informativeness of a set of markers for parentage inference. Previous formulas invariably assume that parentage is to be sought for a single offspring, while in practice multiple full siblings might be sampled (for example, seeds, eggs or young from a pair of monogamous parents) and their father, mother or both are to be assigned among a number of candidates. In this study, I derive formulas for parentage exclusion probabilities for an arbitrary number (n) of fullsibs, which reduce to previous equations for the special case of $n = 1$. I also derive sibship exclusion probabilities, and investigate the power of differentiating half-sib, avuncular and grandparent–grandoff-spring relationships using unlinked autosomal markers among different numbers of tested individuals. Applications of the formulas are demonstrated using both theoretical and empirical data sets of allele frequencies. The results from the study highlight the conclusion that the power of genealogical relationship inferences can be enhanced enormously by analysing multiple individuals for a given set of markers. The equations derived in this study allow more accurate determination of marker information and of the power of a parentage/sibship analysis. In addition, they can be used to guide experimental designs of parentage analyses in selecting markers and determining the number of offspring to be sampled and genotyped.

Heredity (2007) **99**, 205–217; doi:10.1038/sj.hdy.6800984; published online 9 May 2007

Introduction

Parentage inferences from molecular marker data have been widely applied to the studies of social behaviour/organization, reproductive success, mating systems, dispersal and spatial genetic structure in natural populations (Hughes, 1998; Coltman *et al.*, 1999; Garant *et al.*, 2001; Avise *et al.*, 2002; Robledo-Arnuncio and Gil, 2005; Bretman and Tregenza, 2005). Several classes of statistical methods are proposed to perform parentage analyses using data of various genetic markers (Marshall *et al.*, 1998; Jones and Ardren, 2003). To evaluate the statistical power of a parentage analysis and characterize the informativeness of markers, the parentage exclusion probability (P_E) is usually calculated. It is defined as the average capability of any marker system to exclude a 'random' individual from parentage when the other parent (its genotype) is either known or unknown, or to exclude a 'random' pair of individuals as both parents of an offspring. A high P_E value indicates that the marker system is highly informative for parentage analysis and that the parentage analysis using the marker system is highly powerful.

P_E calculation was first described by Wiener *et al.* (1930) for biallelic loci, and was subsequently extended to loci with any number of codominant alleles (Jamieson,

1965; Selvin, 1980; Ohno *et al.*, 1982; Chakraborty *et al.*, 1988; Dodds *et al.*, 1996; Weir 1996, p 209; Jamieson and Taylor, 1997) and to dominant loci (Chakraborty *et al.*, 1974; Gerber *et al.*, 2000). The probability of excluding a relative of (rather than a random individual unrelated to) the true father from paternity when the maternal genotype is known was also derived (Salmon and Brocteur, 1978; Thompson and Meagher, 1987; Double *et al.*, 1997; Fung *et al.*, 2002; Hu *et al.*, 2005).

In all previous studies, however, P_E is calculated invariably assuming that a single offspring is genotyped to infer its parent or pair of parents. A single offspring genotype contains only one paternal and one maternal allele at an autosomal diploid locus, and has no information about the other paternal and maternal alleles. The probability that at least one copy of each parental allele is represented in a set n offspring is $1 - 2^{1-n}$, and the potential of the parental genotype being fully inferable increases rapidly with n. Therefore, genotyping multiple offspring increases parentage exclusion probability for any given marker system. Some statistical methods have been developed to use the genotypes of multiple offspring in inferring their common parentage (Emery *et al.*, 2001; Jones, 2001; Sieberts *et al.*, 2002). Calculating P_E for multiple offspring helps in determining more accurately the power of a parentage analysis, in screening markers by their informativeness, and in deciding on the appropriate numbers of offspring and markers to be genotyped in designing a parentage assignment experiment.

Recently, various statistical methods have also been proposed to infer sibships in a one-generation sample of

individuals, using the genotypes of these individuals at a number of marker loci (Painter, 1997; Almudevar and Field, 1999; Thomas and Hill, 2000, 2002; Smith et al., 2001; Beyer and May, 2003; Wang, 2004; Butler et al., 2004). These methods allow the inference of sibships consisting of an arbitrary number of individuals by exclusion (Smith et al., 2001; Butler et al., 2004) or likelihood (Thomas and Hill, 2000; Wang, 2004) approaches. Although two individuals cannot be excluded from a full sibship using any number of autosomal markers, three or more non-siblings can be excluded from a full sibship using autosomal markers with three or more codominant alleles (Almudevar and Field, 1999; see below). Sibship exclusion probability (S_E) can be defined as the average capability of any marker system to exclude a group of 'random'-unrelated individuals from a full-sib family. Similar to P_E, S_E can be calculated to evaluate the informativeness of markers in and the power of a sibship analysis. A formula for S_E was derived by Almudevar and Field (1999).

In this paper, I derive equations for parentage and sibship exclusion probabilities when an arbitrary number of individuals are involved. The equations for parentage exclusion probabilities reduce to previous ones when only one offspring is genotyped and used in parentage analysis. The equations for sibship exclusion probabilities are more explicit and easier to calculate than previous ones (Almudevar and Field, 1999). I show that, for both parentage and sibship inferences, the power of analysis and amount of marker information increase rapidly with an increasing number of individuals involved in the exclusion analysis. Finally, I consider the inference of half-sib (HS), grandparent–grandchild and avuncular relationships, which have the same IBD (identity by descent) sharing between pairs of individuals and are thus indistinguishable using unlinked autosomal markers (Epstein et al., 2000; McPeek and Sun, 2000). I show that when one or more full siblings of each of the two individuals are also genotyped for some unlinked autosomal markers and included in the relationship analysis, however, the three relationships can be easily discriminated with a high statistical power.

Assumptions

I consider the use of autosomal diploid markers with an arbitrary number of codominant alleles in parentage or sibship analyses. The markers are assumed to follow Mendelian inheritance without mutations and genotyping errors, to be in Hardy–Weinberg equilibrium and to be unlinked and in linkage equilibrium. The allele frequencies of a marker are known. These assumptions were also made explicitly or implicitly in previous calculations of P_E.

Parentage exclusion probability

In classical parentage analyses, an individual is excluded as the parent of an offspring if the offspring's genotype at some locus cannot be generated from the genotype of the individual as a parent following Mendelian inheritance and barring mutations. The average probability (P_E) that a 'random' individual unrelated to the true parent of an offspring is excluded from the parentage of the offspring using the information from a marker can be calculated, which depends on the allele frequencies of the marker

only and indicates the informativeness (capability) of the marker in a parentage analysis. P_E also measures the power of a parentage analysis using a given set of markers.

Three cases of parentage exclusion can be distinguished in practice. An individual is excluded from parentage of an offspring when the other parent (its genotype) is either known or unknown, or a pair of individuals is excluded as both parents of an offspring. Traditionally, P_E is calculated for the three cases assuming a single offspring is genotyped to infer its parentage. I extend previous studies by considering an arbitrary number of full-sib offspring being genotyped to infer their parentage.

Parentage exclusion probability when one parent is known, P_{E1}: Without loss of generality, I suppose a number of n ($\geqslant 1$) full-sib offspring and their mother are genotyped at an autosomal locus with k codominant alleles (A_u) of known frequencies (p_u, $u = 1, 2, ..., k$). The problem is to obtain the probability that a random male unrelated to the true parents is excluded from paternity of the n offspring, using the offspring and mother genotypes and the allele frequencies of the marker. Table 1 lists all possible mother–father–offspring genotype combinations and the corresponding excluded paternal genotypes given mother and offspring genotypes. It also shows, for each combination, the probabilities of the mother genotype, father genotype, offspring genotypes (conditional on parents' genotypes), and the excluded paternal genotype given the mother's and offspring's genotypes. P_{E1} is obtained by summing the product of the joint probability of a mother–father–offspring genotype combination and the corresponding exclusion probability. For illustration, consider row 2 of Table 1 as an example. The joint probability of an A_uA_u mother (with probability p_u^2, column 2), an A_vA_y ($v \neq y$) father (with probability $2p_up_y$, column 4) and n A_uA_v offspring (with probability $(1/2)^n$ conditional on parental genotypes, column 6) is $p_u^2 \times 2p_vp_y \times (\frac{1}{2})^n$, and this mother–offspring genotype combination excludes all males that do not have an A_v allele. Such males have a frequency of $(1-p_v)^2$ (column 8), so that the combination listed in row 2 of Table 1 has a combined exclusion probability of $p_u^2 \times 2p_vp_y \times (\frac{1}{2})^n \times (1 - p_v)^2$. Adding these probabilities for all mother–father–offspring genotype combinations listed in the table leads, after some tedious algebra, to

$$
\begin{aligned}
P_{E1} = {} & 1 - 4ca_2 - 2(1 + 4b - 2c)a_2^2 + 8(b + c - d)a_2^3 \\
& - 2(1 - 3c)a_3 - 2(6b + 3c - 4d)a_3^2 \\
& + (3 + 8b - 6c)a_4 - 4(9b + 6c - 7d)a_2a_4 \\
& + 4(7b + c - 2d)(a_2a_3 - a_5) + 2(20b + 11c - 14d)a_6
\end{aligned}
\tag{1}
$$

where

$$ b = (\tfrac{1}{4})^n, c = (\tfrac{2}{4})^n, d = (\tfrac{3}{4})^n $$

and $a_s = \sum_{u=1}^{k} p_u^s$, the sum of the sth power of allele frequencies.

For a single offspring ($n = 1$), (1) simplifies to

$$
P_{E1} = 1 - 2a_2 - 2a_2^2 + a_3 + 2a_4 - 3a_5 + 3a_2a_3,
\tag{2}
$$

which is the same as derived earlier (Jamieson and Taylor, 1997). For a given number of offspring (n) and a given

Table 1 Paternity exclusion configurations of multiple offspring genotypes at a k-codominant allele locus: mother known

Mother		Father		Offspring		Excluded male	
Type	Probability[a]	Type	Probability[b]	Genotype[c]	Probability[d]	Genotype $= A_wA_x$	Probability[e]
A_uA_u	p_u^2	A_vA_v	p_v^2	A_uA_v	1	$w,x \neq v$	$(1-p_v)^2$
		$A_vA_y\ v \neq y$	$2p_vp_y$	A_uA_v	$(1/2)^n$	$w,x \neq v$	$(1-p_v)^2$
				A_uA_y	$(1/2)^n$	$w,x \neq y$	$(1-p_y)^2$
				A_uA_v,A_uA_y	$1-(1/2)^{n-1}$	$w,x \neq v,y$	$(1-2p_vp_y)^2$
$A_uA_v\ v \neq u$	$2p_up_v$	$A_tA_t\ t=u,v$	p_t^2	A_uA_v	$(1/2)^n$	$w,x \neq u;\ w,x \neq v$	$(1-p_u-p_v)^2$
				$A_tA_t;\ A_tA_t,A_uA_v$	$1-(1/2)^n$	$w,x \neq t$	$(1-p_t)^2$
		A_uA_v	$2p_up_v$	A_uA_v	$(1/2)^n$	$w,x \neq u;\ w,x \neq v$	$(1-p_u-p_v)^2$
				$A_tA_t,A_uA_v;\ A_tA_t\ (t=u,v)$	$(3/4)^n-(1/2)^n$	$w,x \neq t$	$(1-p_t)^2$
				A_uA_u,A_vA_v,A_uA_v	$1+(1/2)^n-2(3/4)^n$	$w,x \neq u,v$	$1-2p_up_v$
		$A_tA_y\ y \neq u,v;\ t=u,v$	$2p_tp_y$	$A_tA_t;\ A_tA_t,A_uA_v$	$(1/2)^n-(1/4)^n$	$w,x \neq t$	$(1-p_t)^2$
				A_uA_v	$(1/4)^n$	$w,x \neq u;\ w,x \neq v$	$(1-p_u-p_v)^2$
				$A_uA_y;\ A_vA_y;\ A_uA_yA_vA_y$	$(1/2)^n$	$w,x \neq y$	$(1-p_y)^2$
				$A_uA_v,A_uA_y;\ A_uA_v,A_vA_y;$ A_uA_v,A_uA_y,A_vA_y	$(3/4)^n-(1/2)^n-(1/4)^n$	$w,x \neq u,y;$	$1-2(p_u+p_v)p_y$
						$w,x \neq v,y$	
				Other	$1-(3/4)^n-(1/2)^{n+}(1/4)^n$	$w,x \neq t,y$	$1-2p_tp_y$
		$A_yA_y\ y \neq u,v$	p_y^2	$A_uA_y;\ A_vA_y;\ A_uA_yA_vA_y$	1	$w,x \neq y$	$(1-p_y)^2$
		$A_yA_z\ y \neq u,v;\ z \neq u,v;$ $y \neq z$	$2p_yp_z$	$A_uA_t;\ A_vA_t;\ A_uA_t,A_vA_t$ $(t=y,z)$	$(1/2)^n$	$w,x \neq t$	$(1-p_t)^2$
				Other	$1-(1/2)^{n-1}$	$w,x \neq y,z$	$1-2p_yp_z$

[a]Probability of mother's genotype.
[b]Probability of father's genotype.
[c]Different combinations of offspring genotypes are separated by semicolons, and 'Other' refers to all other offspring genotype combinations given mother's and father's genotypes. Within a combination, genotypes are separated by commas. Some genotype combinations are denoted by a generic allele index t, defined in brackets.
[d]Probability of n offspring's genotypes given mother's and father's genotypes.
[e]Probability of excluded genotype given mother's and offspring's genotypes.

number of alleles (k) at a locus, P_{E1} increases as the allele frequencies become increasingly even. The maximum value is reached when all k alleles have an equal frequency of $1/k$,

$$
\begin{aligned}
P_{E1} =& 1 - (2^n k^4 + (4^n + 2 - 5 \times 2^{n-1}) k^3 \\
& - (11 + 3 \times 2^{n-1} - 4 \times 3^n + 3 \times 4^{n-1}) k^2 \\
& + (19 + 17 \times 2^{n-1} - 11 \times 3^n) k \\
& - (10 + 11 \times 2^{n-1} - 7 \times 3^n)) / (4^{n-1} k^5)
\end{aligned}
\tag{3}
$$

which again reduces to

$$
P_{E1} = 1 - (2k^3 + k^2 - 5k + 3)/k^4
\tag{4}
$$

when $n = 1$ as derived previously (Weir, 1996). P_{E1} is also a monotonically increasing function of the number of offspring (n), and the maximum value when $n \to \infty$ is

$$
P_{E1} = 1 - 2a_2^2 - 2a_3 + 3a_4
\tag{5}
$$

The maximum value computed by (5) is generally quickly attained with an increasing n, except when the marker is very uninformative (that is, few alleles with uneven frequencies).

Parentage exclusion probability when no parent is known, P_{E2}:

An individual can be excluded from the parentage of a group of offspring if the genotype of any offspring at a locus cannot be generated from the genotype of the individual. Without loss of generality, I assume a number of n ($\geqslant 1$) full-sib offspring are genotyped at an autosomal locus with k codominant alleles to infer their paternity without knowledge of their maternal genotype. The (average) probability that a random male unrelated to the true parents is excluded from paternity of the n offspring, using the offspring genotypes and allele frequencies of the marker, can be derived similar to P_{E1},

$$
\begin{aligned}
P_{E2} =& 1 - 8ca_2 - 4(1 - 6b)a_2^2 - 8(3b - 3c + d)a_2^3 \\
& - 4(1 - 6b - c)a_3 - 8(21b - 12c + d)a_2a_3 \\
& + 2(3 + 48b - 36c + 4d)a_3^2 + 6(1 - 14b + 4c)a_4 \\
& + 4(2 + 45b - 37c + 7d)a_2a_4 \\
& + 2(1 + 108b - 62c + 4d)a_5 \\
& - (15 + 264b - 204c + 28d)a_6.
\end{aligned}
\tag{6}
$$

where a, b, c and d are as defined in (1). As is expected, (6) reduces, for a single offspring ($n = 1$), to

$$
P_{E2} = 1 - 4a_2 + 2a_2^2 + 4a_3 - 3a_4
\tag{7}
$$

as derived previously (Jamieson and Taylor, 1997). Similar to P_{E1}, P_{E2} increases as the allele frequencies become increasingly even for a given number of offspring (n) and a given number of alleles at a locus. The maximum value is reached when all k alleles have frequency $1/k$,

$$
\begin{aligned}
P_{E2} =& 1 - (2^{n+1} k^4 - 2(6 + 2^{n-1} - 4^n) k^3 \\
& + (69 - 9 \times 2^{n+2} + 4 \times 3^n - 6 \times 4^{n-1}) k^2 \\
& - (123 - 86 \times 2^n + 11 \times 3^n + 4^{n+1}) k \\
& + 3(22 - 17 \times 2^n + 7 \times 3^{n-1} + 5 \times 4^{n-1})) / (4^{n-1} k^5)
\end{aligned}
\tag{8}
$$

When $n = 1$, (8) reduces to

$$
P_{E2} = 1 - (4k^2 - 6k + 3)/k^3
\tag{9}
$$

as derived previously (Jamieson and Taylor, 1997). Like P_{E1}, P_{E2} is also a monotonically increasing function of the number of offspring (n), and the maximum value when $n \to \infty$ is

$$
\begin{aligned}
P_{E2} =& 1 - 4a_2^2 - 4a_3 + 6a_3^2 + 6a_4 + 8a_2a_4 + 2a_5 \\
& - 15a_6
\end{aligned}
\tag{10}
$$

Exclusion probability of a pair of individuals as parents, P_{E3}

For a group of n ($\geqslant 1$) full-sib offspring, we may be interested in inferring the pair of parents that have produced them. A pair of individuals can be excluded as both parents of the n offspring if their genotypes cannot be explained fully by those of the two individuals as parents at a locus. The average probability of excluding two random individuals, who are unrelated between themselves and to the true parents, as parents of the n offspring using an autosomal marker with k codominant alleles is

$$
\begin{aligned}
P_{E3} =& 1 - 32ba_2^2 - 8(1 - 2c)a_2^4 - 16(1 - 2b - c)a_2^2a_3 \\
& - 8(1 + 3b - 4c)a_3^2 - 64(3b - 3c + d)a_2a_3^2 \\
& + 16ba_4 + 8(1 - 24b + 10c)a_2^2a_4 \\
& + 8(3 - 26b + 13c - 4d)a_3a_4 \\
& + 2(5 + 120b - 94c + 16d)a_4^2 \\
& - 16(3b - c)(a_5 - 2a_2a_3) \\
& - 16(2b - c)a_2(3a_4 + 8a_5 - 2a_2^2) \\
& + 16(1 + 31b - 31c + 9d)a_3a_5 \\
& + 4(1 + b - 4c)a_6 \\
& + 32(1 + 20b - 15c + 2d)a_2a_6 \\
& - 4(1 - 114b + 67c - 8d)a_7 \\
& - (996b - 880c + 176d + 59)a_8
\end{aligned}
\tag{11}
$$

which can be derived using an approach similar to that in deriving P_{E1} and P_{E2}. In (11), a, b, c and d are as defined in (1). For the case of a single offspring ($n = 1$), (11) is simplified to

$$
P_{E3} = 1 - 8a_2^2 + 8a_2a_3 + 2a_3^2 + 4a_4 - 4a_5 - 3a_6
\tag{12}
$$

as derived earlier (Jamieson and Taylor, 1997). Similar to P_{E1} and P_{E2}, the maximum value of P_{E3} is reached when all k alleles at a locus have frequency $1/k$,

$$
\begin{aligned}
P_{E3} =& 1 - (8k^5 - 4(11 - 2^{n+2}) k^4 + 2(17 - 2^{n+4} + 4^{n+1}) k^3 \\
& + (211 - 61 \times 2^{n+1} + 24 \times 3^n - 9 \times 4^n) k^2 \\
& - 2(229 - 179 \times 2^n + 34 \times 3^n + 27 \times 4^{n-1}) k \\
& + 249 - 55 \times 2^{n+2} + 44 \times 3^n + 59 \times 4^{n-1}) / (4^{n-1} k^7)
\end{aligned}
\tag{13}
$$

When $n = 1$, (13) reduces to

$$P_{E3} = 1 - (8k^3 - 12k^2 + 2k + 3)/k^5 \qquad (14)$$

as derived earlier (Jamieson and Taylor, 1997). Like P_{E1} and P_{E2}, P_{E3} is also a monotonically increasing function of the number of offspring (n), and its maximum value when $n \to \infty$ is

$$\begin{aligned} P_{E3} = {}&1 - 8a_2^4 - 16a_2^2 a_3 - 8a_3^2 + 8a_2^2 a_4 + 24a_3 a_4 \\ &+ 10a_4^2 + 4a_6 + 32a_2 a_6 - 4a_7 + 16a_3 a_5 - 59a_8 \end{aligned}$$

$$(15)$$

Sibship exclusion probability

A group of individuals are excluded from comprising a full sibship if their genotypes at a locus cannot be generated by any pair of parental genotypes. For an autosomal codominant locus, sibship exclusion is warranted, for example, when the individuals display five or more alleles. Like parentage exclusion, we can calculate the average probability, S_E, of excluding a group of n-unrelated individuals as full siblings using a k-allele codominant marker. Therefore, S_E signifies the capability of a marker in a sibship analysis, and the statistical power of a sibship analysis using a given set of markers.

For an autosomal diploid locus with k codominant alleles, the average probability of excluding n-unrelated individuals as full siblings can be derived (Appendix) as

$$\begin{aligned} S_E = {}&1 - \tfrac{1}{2}(k-2)(k-3) \\ &\times \sum_u p_u^{2n} - \tfrac{1}{2} \sum_u \sum_{v \neq u} b_{uv}^{2n} - \tfrac{1}{6} \\ &\times \sum_u \sum_{v \neq u} \sum_{w \neq u, v} (3(p_u c_{uv} + 2p_w b_{uv})^n \\ &\quad - 2(p_u c_{vw} + p_v c_{uw})^n \\ &\quad - 3p_u^n (c_{uv}^n + c_{uw}^n) \\ &\quad - 2^n (p_w^n b_{uv}^n + p_v^n b_{uw}^n - 2p_u^n b_{vw}^n - d_{vw}^n - d_{uv}^n - d_{uw}^n)) \\ &- 2^{n-3} \sum_u \sum_{v \neq u} \sum_{w \neq u, v} \sum_{x \neq u, v, w} (2d_{uv}^n + 3d_{uw}^n - d_{vx}^n \\ &\quad + 2d_{wx}^n + (p_x^n - 3p_u^n + b_{ux}^n) b_{vw}^n + (p_v^n - 3p_w^n) b_{ux}^n \\ &\quad + (d_{uw} + d_{vx})^n - (d_{uv} + d_{uw} + d_{vx})^n - 2(d_{uv} + d_{wx})^n \\ &\quad + 3(d_{uv} + d_{uw} + d_{wx})^n - (d_{uv} + d_{vx} + d_{wx})^n \\ &\quad - (d_{uw} + d_{vx} + d_{wx})^n) \end{aligned}$$

$$(16)$$

where $b_{st} = p_s + p_t$, $c_{st} = p_s + 2p_t$, $d_{st} = p_s p_t$ (note $c_{st} \neq c_{ts}$) for $s, t = u, v, w, x = 1, \ldots, k$. S_E simplifies greatly in the following special cases.

(1) $n = 2$ or $k = 2$

It can be shown that $S_E \equiv 0$ when either $n = 2$ or $k = 2$. A pair of individuals ($n = 2$) is never excluded from being full siblings no matter how polymorphic a marker is, and biallelic loci ($k = 2$) do not allow sibship exclusion regardless of n.

(2) $n = 3$ and $n = 4$

For trios and quadruplets of unrelated individuals, (16) reduces to

$$\begin{aligned} S_E = {}&1 - 30a_2^2 + 16a_2^3 - 22a_3^2 + 72a_2 a_3 \\ &- 60a_2 a_4 + 15a_4 - 48a_5 + 56a_6 \end{aligned}$$

$$(17)$$

$$\begin{aligned} S_E = {}&1 - 60a_2^4 - 192a_2^2 a_3 - 112a_3^2 + 288a_2 a_3^2 \\ &+ 396a_2^2 a_4 + 480a_3 a_4 - 315a_4^2 + 240a_2 a_5 \\ &- 552a_3 a_5 + 84a_6 - 696a_2 a_6 - 480a_7 + 918a_8 \end{aligned}$$

$$(18)$$

respectively,
where

$$a_s = \sum_{u=1}^{k} p_u^s$$

(3) Equal allele frequency

Like P_E, S_E increases for given values of n (> 2) and k (> 2) when allele frequencies become increasingly even. The maximum S_E is attained when all alleles have the same frequency of $1/k$,

$$\begin{aligned} S_E = {}&1 - (3 - 7 \times 2^{n-1} - 2 \times 3^n + 13 \times 4^{n-1} - 4 \times 6^{n-1} \\ &+ 7^n - 6 \times 8^{n-1}) k^{1-2n} + 2^{-1} (5 - 27 \times 2^{n-1} \\ &- 6 \times 3^n + 51 \times 4^{n-1} - 2 \times 6^n + 3 \times 7^n \\ &- 22 \times 8^{n-1}) k^{2-2n} - 2^{-2} (2 - 2^{n+4} - 4 \times 3^n \\ &+ 15 \times 4^n - 8 \times 6^{n-1} + 2 \times 7^n - 3 \times 8^n) k^{3-2n} \\ &- 2^{n-3} (6 - 5 \times 2^n + 4^n) k^{4-2n} \end{aligned}$$

$$(19)$$

Exclusion probabilities for multiple loci and multiple tests

For a number of L-independent loci, the cumulative exclusion probability is calculated as

$$P = 1 - \prod_{l=1}^{L} (1 - P_l) \qquad (20)$$

where P_l is the exclusion probability for locus l calculated by (1), (6), (11) or (16).

The above calculations are for a single test. In almost all practical analyses, however, usually a large number of groups of individuals are tested for parentage or sibship and the aim is ideally to exclude all false parentage or sibship relationships. For a given marker system, the aim is obviously more difficult to achieve with a larger number of tests. For a number of M independently replicated tests, the number of non-exclusions of a false relationship, m, is roughly binomially distributed, $m \sim$ Binomial ($M, 1 - P$), with a mean of $M(1 - P)$ and a variance of $MP(1 - P)$. Using a given set of markers, the number of non-excluded false parentage (or sibship) events is expected to increase linearly with the number of tests M. The probability that exclusions occur to all of the M tests (that is, perfect exclusion of all false relationships) is P^M, which decreases exponentially with M.

Theoretical examples of exclusion probabilities

Parentage (sibship) exclusion probabilities depend on the allele frequencies at a locus and the number of full-sib offspring (the number of unrelated individuals) that are genotyped for determining their parentage (sibship), n. The effects of allele frequencies on exclusion probabilities
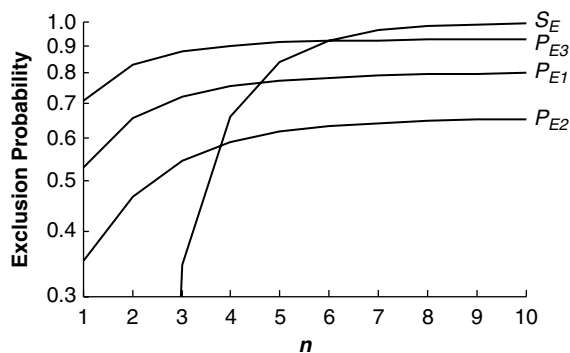
Figure 1 Exclusion probabilities as a function of the number (n) of full-sib offspring in a parentage analysis or of unrelated individuals in a sibship analysis. A single locus having five codominant alleles with frequencies in a triangular distribution is used in the calculation.
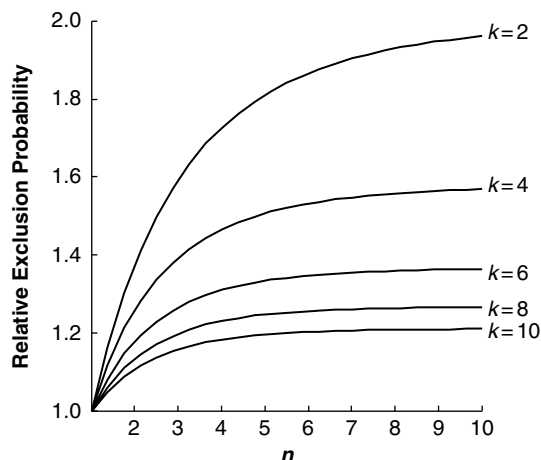

Figure 2 Paternity exclusion probabilities for multiple offspring ($n > 1$) relative to those for a single offspring ($n = 1$). The maternal genotypes are assumed known, and a locus with k alleles of an equal frequency is used in calculating P_{E1}. The five lines correspond to $k = 2, 4, 6, 8, 10$ as indicated in the graph.

are well known. As a numerical example to illustrate the effect of n, I calculated exclusion probabilities using a locus with k alleles in triangular frequencies of $p_i = i/(k(k+1)/2)$ for $i = 1, \ldots, k$. The changes of P_{E1}, P_{E2}, P_{E3} and S_E with n for a locus with $k = 5$ alleles are shown in Figure 1. For any given value of n, P_{E3} is the largest and P_{E2} is the smallest among the three parentage exclusion probabilities. P_{E1} is always larger than P_{E2} because the extra maternal information allows more exclusions of false fathers. All of the four exclusion probabilities increase with an increasing value of n and quickly become attenuated at $n = 4$–7. For this particular locus, the minimum values of P_{E1}, P_{E2}, P_{E3} and S_E are 0.53, 0.35, 0.71 and 0.00, respectively, when $n = 1$, and maximum values are 0.81, 0.66, 0.93 and 1.00, respectively, when $n > 7$.

Increasing n is more beneficial for less informative markers. Figure 2 plots exclusion probabilities (P_{E1}) for multiple offspring ($n > 1$) relative to that for a single offspring ($n = 1$) as a function of n. A locus with k ($= 2, 4, 6, 8, 10$) equally frequent alleles was used in calculating P_{E1} from (3). As can be seen, a smaller k leads to a faster increase in P_{E1} with n. While the relative exclusion probability for $n = 10$ is only about 120% for a highly informative locus with $k = 10$ alleles, it is about 200% for a much less informative biallelic locus ($k = 2$). Therefore, less informative markers are more efficiently compensated by sampling and genotyping multiple offspring in parentage analyses. The same conclusion is true for P_{E2} and P_{E3}, and for loci with any allele frequency distributions.

Because parentage exclusion probabilities increase with both the number of loci and the number of full-sib offspring, one may want to know in a practical parentage analysis whether it is more rewarding to genotype more loci or more offspring for a given cost or alternatively whether it is more economical to genotype more loci or more offspring to achieve a given statistical power. Figure 3 shows the exclusion probabilities for $L = 1$ and $n = 2, 4, 6$ relative to those for $L = 2$ and $n = 1$, as a function of the number of alleles (k) per locus. The allele frequencies are assumed to be in a triangular distribution. As can be seen, $L = 1$ and $n = 6$ results in similar (for P_{E1} and P_{E3}) or even larger (for P_{E2}) parentage exclusion
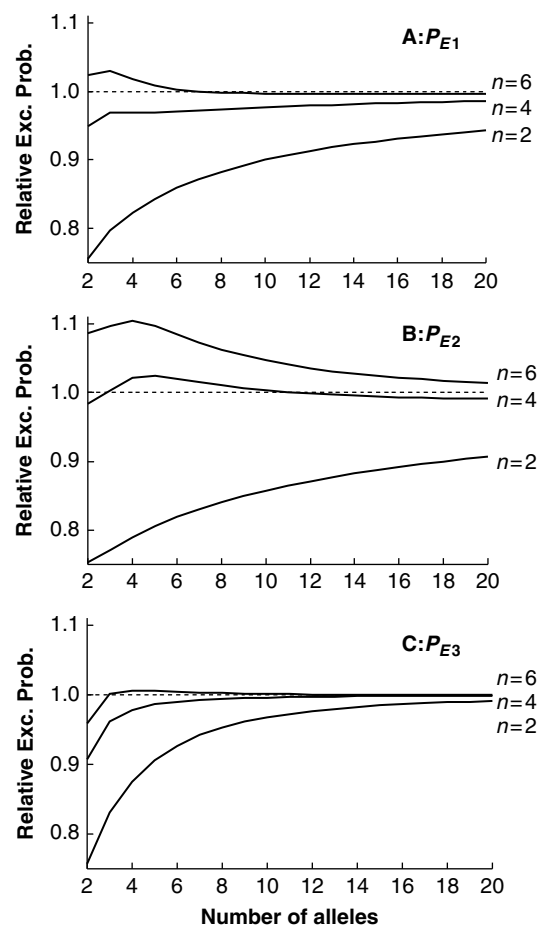

Figure 3 Parentage exclusion probabilities with a single locus and multiple offspring relative to those with two loci and a single offspring. It is assumed that a locus has 2–20 codominant alleles with frequencies in a triangular distribution, and the number (n) of full-sib offspring used in parentage analyses is 2, 4 or 6. The upper, middle and lower panels are for relative parentage exclusion probabilities P_{E1}, P_{E2} and P_{E3}, respectively.

probabilities than $L=2$ and $n=1$. While one should always try to use as many markers as possible in a parentage analysis for maximal power, he/she can also improve the analysis power by genotyping more offspring. The development of markers such as microsatellites in relationship analyses for a given species is expensive, and furthermore, the number of informative markers available may be limited. In such situations, therefore, the power of parentage analyses can be improved substantially by genotyping more offspring in a litter.

Practical examples of exclusion probabilities

The allele frequencies of seven microsatellite loci in Atlantic salmon were published in Villanueva et al. (2002). The parentage and sibship exclusion probabilities for the loci are calculated using (1), (6), (11) and (16) and are listed in Table 2. The most informative locus for parentage or sibship inference is locus 1 which has 14 alleles, rather than locus 2 that has 21 alleles. This is because the allele frequencies of locus 1 are more even than those of locus 2. The rank order of the informativeness among the seven markers changes only slightly, depending on the specific exclusion probability being computed and compared. Locus 2 gives slightly larger values of P_{E1} and P_{E2} but smaller values of P_{E3} and S_E than locus 3. For each of the seven loci, its exclusion power increases rapidly with an increasing number of full-sib offspring genotyped and included in a relationship analysis.

The combined exclusion probabilities over the seven loci are very high, close to the maximum value of 1. However, this does not necessarily mean that all false relationships can be excluded in a given parentage analysis. The accuracy of a parentage analysis depends on not only how informative the markers are as indicated by exclusion probabilities, but also the total number of tests to be carried out in the analysis. As an example, consider the exclusion of paternity for offspring with known maternal genotypes. The combined P_{E1} for the seven microsatellites is $1-2.00 \times 10^{-4} = 0.9998$ when a single offspring ($n=1$) is tested for paternity. If a sample of N offspring–mother pairs and N candidate fathers unrelated to any of the N offspring is obtained, and each candidate is tested for paternity of each offspring, there would be a total number of N^2 tests. These N^2 tests are not truly independent, because an individual appears in multiple tests. However, to a good approximation, the following calculation is conducted under the assumption of independent tests. The probability of complete exclusion of the N^2 false offspring–father relationships is 0.9998^{N^2}, which is 0.923 for $N=20$, 0.135 for $N=100$ and 0.000335 for $N=200$. With an increasing sample size, the exclusion power of the 7 microsatellites diminishes exponentially. If there are four full-sib offspring instead of only one offspring in each of the N mother–offspring group, the combined P_{E1} for the seven microsatellites is increased to $1-3.13 \times 10^{-7} = 0.999999687$. The probability of complete exclusion of the N^2 false offspring–father relationships becomes 0.999999687^{N^2}, which is 0.9999 for $N=20$, 0.9969 for $N=100$ and 0.9876 for $N=200$.

The above numerical examples illustrate that a high exclusion probability may still result in a low probability

Table 2 Exclusion probabilities calculated for seven microsatellite loci in Atlantic salmon[a]

Locus/k	Allele frequency	n=1			n=2			n=4			
		P_{E1}	P_{E2}	P_{E3}	P_{E1}	P_{E2}	P_{E3}	P_{E1}	P_{E2}	P_{E3}	S_E
1/14	$2 \times 0.167, 0.109, 2 \times 0.090, 0.083, 0.077,$ $0.064, 0.045, 2 \times 0.032, 2 \times 0.019, 0.006$	0.787	0.648	0.929	0.872	0.763	0.972	0.934	0.872	0.992	0.972
2/21	$0.279, 0.100, 0.093, 2 \times 0.087, 0.067, 0.053,$ $0.040, 3 \times 0.033, 0.020, 2 \times 0.013, 7 \times 0.007$	0.762	0.614	0.919	0.849	0.725	0.966	0.913	0.836	0.988	0.961
3/15	$0.204, 0.197, 0.102, 0.080, 2 \times 0.073, 0.066,$ $0.051, 2 \times 0.044, 0.022, 2 \times 0.015, 2 \times 0.007$	0.763	0.616	0.915	0.853	0.732	0.965	0.919	0.846	0.988	0.959
4/13	$0.212, 0.159, 0.141, 0.135, 0.100, 0.082,$ $0.053, 0.035, 0.029, 0.024, 0.018, 2 \times 0.006$	0.738	0.582	0.897	0.836	0.704	0.956	0.907	0.825	0.985	0.943
5/12	$0.226, 0.171, 0.140, 0.128, 0.116, 0.110,$ $0.043, 0.031, 2 \times 0.012, 0.006, 0.005$	0.709	0.546	0.875	0.814	0.672	0.944	0.891	0.798	0.979	0.921
6/7	$0.378, 0.209, 0.136, 2 \times 0.096, 0.079, 0.006$	0.571	0.390	0.761	0.692	0.505	0.867	0.784	0.633	0.929	0.756
7/9	$0.524, 0.141, 0.080, 2 \times 0.067, 0.054, 0.047,$ $0.013, 0.007$	0.492	0.303	0.706	0.607	0.393	0.816	0.694	0.503	0.883	0.649
All Loci[b]		$1-2.00$ $\times 10^{-4}$	$1-4.21$ $\times 10^{-3}$	$1-4.42$ $\times 10^{-7}$	$1-1.05$ $\times 10^{-5}$	$1-5.10$ $\times 10^{-4}$	$1-1.97$ $\times 10^{-9}$	$1-3.13$ $\times 10^{-7}$	$1-2.08$ $\times 10^{-5}$	$1-2.90$ $\times 10^{-12}$	$1-1.73$ $\times 10^{-8}$

[a]The allele frequency data were published by Villanueva et al. (2002).
[b]The multiple-locus exclusion probabilities, calculated by equation (20), are close to 1 and are thus expressed as $1-x$, where x is non-exclusion probability.

of complete exclusion of all false parentage if the sample size is large, and that the use of multiple full-sib offspring can increase the power dramatically. In a similar context, we should realize the importance of recording exclusion probabilities with a sufficient number of significant digits. In the above numerical example of N offspring–mother pairs ($n = 1$) and N candidate fathers with $N = 100$, the probability of complete exclusion of the 10^4 false offspring–father relationships is 0.135 if $E_{P1} = 0.9998$ but becomes 0.368 if $E_{P1} = 0.9999$ and 0.050 if $E_{P1} = 0.9997$. A tiny change in exclusion probability can translate to a substantial alteration in the measurements of the overall power of a parentage analysis. For this reason, it is more convenient to calculate and record non-exclusion probabilities rather than exclusion probabilities.

Distinguishing half-sib, avuncular and grandparent–grandoffspring relationships

In a parentage analysis, the use of multiple full-sib offspring increases dramatically the probability of excluding a false parent or a false pair of parents as shown above. Analysing trios rather than pairs of individuals simultaneously for genealogical relationships in a likelihood framework also increased the power substantially (Sieberts et al., 2002). It is well known that the three relationships, half-sib (HS), avuncular and grandparent–grandoffspring (GG), between a pair of

individuals cannot be distinguished using unlinked autosomal markers and can be distinguished with very low power using linked autosomal markers (Epstein et al., 2000). In this section, I show that when one or both individuals in the pair have one or more relatives (for example, fullsibs) and the genotype data of the two individuals and their relatives are analysed jointly, these three relationships can be easily differentiated using unlinked autosomal markers. Avuncular refers to any of the four combinations of aunt–nephew, aunt–niece, uncle–nephew and uncle–niece, and I consider aunt–niece (AN) as an example.

Suppose a pair of individuals, A and B, may have a HS, AN or GG relationship, and n_1–1 full siblings to A and n_2–1 full siblings to B are also sampled and genotyped at an autosomal marker with k codominant alleles. Here I consider the likelihood of these $n_1 + n_2$ individuals falling into the three possible pedigrees as depicted in Figure 4. When $n_1 = n_2 = 1$, this reduces to inferring HS, AN and GG relationships between a pair of individuals, A and B. Notice that any pair of individuals taken separately from clusters 1 with n_1 individuals and 2 with n_2 individuals has the same relationship for the HS or AN pedigree, but can have one of two possible relationships, grandparent–grandoffspring or grandaunt–grandniece, for the GG pedigree if $n_1 > 1$ (Figure 4).

The likelihoods of the two full-sib clusters with n_1 and n_2 individuals falling into the HS, AN and GG pedigrees
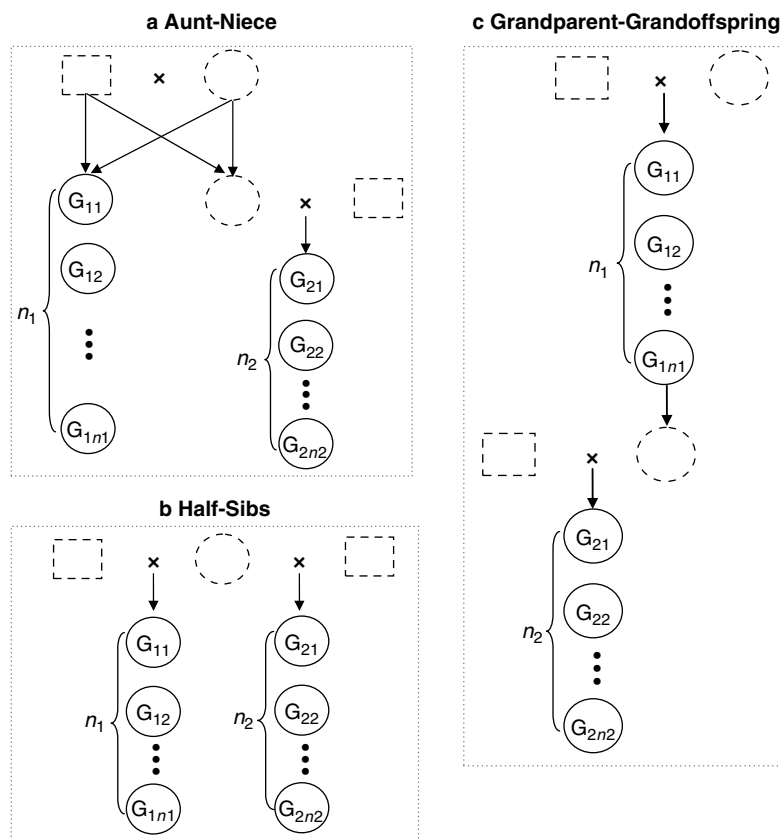

Figure 4 Pedigrees involving aunt–niece (AN), half-sib (HS) and grandparent–grandoffspring (GG) relationships. Males and females are indicated by squares and circles, respectively, and individuals that are sampled and unsampled are indicated by solid and broken lines, respectively.

in Figure 4 are

$$L_{HS} = \sum_{u=1}^{k}\sum_{v=1}^{k} p_u p_v \prod_{i=1}^{2}\left(\sum_{w=1}^{k}\sum_{x=1}^{k} p_w p_x \prod_{j=1}^{n_i}\mathrm{Pr}(G_{ij}|G_{uv},G_{wx})\right)$$

$$L_{AN} = \tfrac{1}{4}\sum_{u=1}^{k}\sum_{v=1}^{k} p_u p_v \sum_{w=1}^{k}\sum_{x=1}^{k} p_w p_x$$
$$\times\left(\prod_{j=1}^{n_1}\mathrm{Pr}(G_{1j}|G_{uv},G_{wx})\right)$$
$$\times\left(\sum_{y=1}^{k}\sum_{z=1}^{k} p_y p_z \sum_{s=u,v}\sum_{t=w,x}\prod_{j=1}^{n_2}\mathrm{Pr}(G_{2j}|G_{st},G_{yz})\right)$$

$$L_{GG} = \tfrac{1}{2}\sum_{u=1}^{k}\sum_{v=1}^{k} p_u p_v \sum_{w=1}^{k}\sum_{x=1}^{k} p_w p_x$$
$$\times\left(\prod_{j=1}^{n_1}\mathrm{Pr}(G_{1j}|G_{uv},G_{wx})\right)$$
$$\times\left(\sum_{y=1}^{k} p_y \sum_{z=1}^{k} p_z \sum_{s=1}^{k} p_s \sum_{t=c,d}\prod_{j=1}^{n_2}\mathrm{Pr}(G_{2j}|G_{yt},G_{zs})\right)$$

where

$$\mathrm{Pr}(G_{ij}|G_{uv},G_{wx}) = \tfrac{1}{4}(\mathrm{Pr}(G_{ij}|G_{uw}) + \mathrm{Pr}(G_{ij}|G_{ux}) + \mathrm{Pr}(G_{ij}|G_{vw}) + \mathrm{Pr}(G_{ij}|G_{vx}))$$

is the probability of observing the genotype of offspring j in cluster i ($i=1,2; j=1,\dots,n_i$), given parental genotypes G_{uv} and G_{wx}. $\mathrm{Pr}(G_{ij}|G_{uw})=1$ if the genotype of offspring j in cluster i has both alleles u and w and $\mathrm{Pr}(G_{ij}|G_{uw})=0$ if otherwise. Note that in L_{GG}, t indexes the two alleles, c and d, in the genotype of the grandparent of full-sib cluster 2.

It can be shown that $L_{HS}\equiv L_{AN}\equiv L_{GG}$ when $n_1=n_2=1$, $L_{HS}\equiv L_{GG}\neq L_{AN}$ when $n_1=1$ and $n_2>1$, $L_{HS}\equiv L_{AN}\neq L_{GG}$ when $n_1>1$ and $n_2=1$, indicating that these three relationships cannot be distinguished no matter how many markers are used when $n_1=1$ and/or $n_2=1$. If both $n_1>1$ and $n_2>1$, however, the three likelihood values are different for an autosomal marker and therefore the three relationships can be differentiated. As a numerical example, the seven microsatellite markers in Atlantic salmons listed in Table 2 are utilized to distinguish the three relationships when the two full-sib clusters have various sizes. The HS, AN or GG pedigrees depicted in Figure 4 are simulated and the genotypes of the two full-sib clusters at the seven microsatellite loci are generated following Mendelian segregation. L_{HS}, L_{AN} and L_{GG} are then calculated from the genotype data, and the relationship between the two clusters of full siblings is inferred as the one with the maximum likelihood. Whenever two or three relationships have the same maximum likelihood, they are assigned as the true relationship with an equal probability. Each pedigree is simulated 100 000 times for a given value of n_1 or n_2, assuming $n_1=n_2$. The rates that an actual relationship is inferred as HS, AN and GG are plotted against n_1 (or n_2) in Figure 5. When $n_1=n_2=1$,
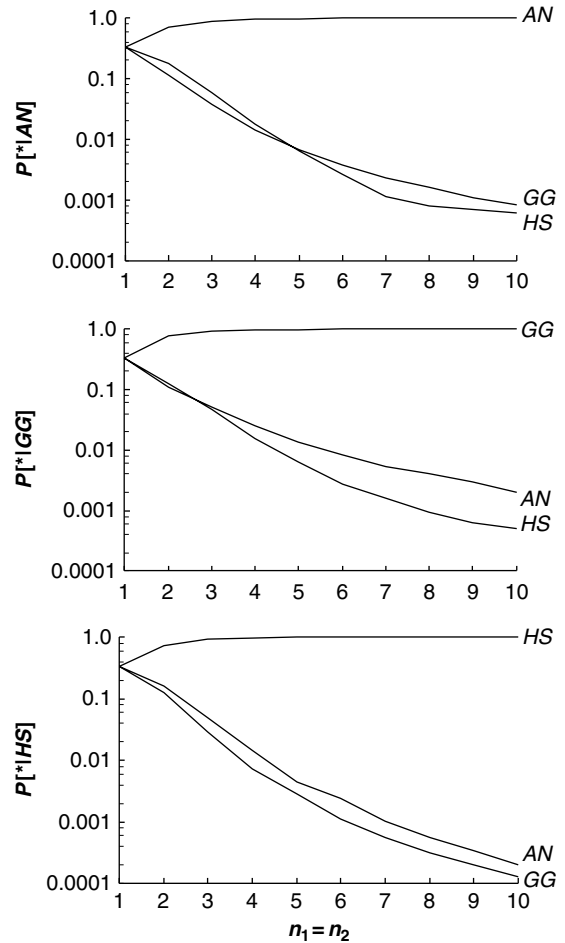

Figure 5 The effect of the number of full siblings on the accuracy of distinguishing aunt–niece, grandparent–grandoffspring and half-sib relationships. Lines marked by AN, GG and HS show the proportions of the actually simulated aunt–niece (top), grandparent–grandoffspring (middle) or half-sib (bottom) relationships between the two full-sib clusters being inferred as AN, GG and HS relationships, respectively. The numbers of individuals in the two full-sib clusters are assumed to be the same ($n_1=n_2$) as shown on the x axis. The data are simulated using the seven microsatellite markers in the Atlantic salmons.

the three relationships are indistinguishable regardless of the actual relationship, resulting in a correct classification rate of 1/3. With an increasing value of n_1 ($=n_2$), however, the relationship-misclassification rate decreases rapidly for all of the three simulated pedigrees. Even with $n_1=n_2=3$, the misclassification rate is only 0.10, 0.10 and 0.08 for the simulated relationship of AN, HS and GG, respectively. The statistical power of the analysis using merely seven microsatellites is extremely high compared with that of the analysis using pairs of individuals but hundreds of linked markers (Epstein et al., 2000). Using 399 autosomal markers (each with four equally frequent alleles) spaced at 10-cM intervals across the human genome, a pair of individuals with GG, HS and AN relationships is still assigned an incorrect relationship at a rate of 0.28, 0.63 and 0.38, respectively (Epstein et al., 2000). The comparison highlights the impact of analysing simultaneously multiple individuals with partially known or completely unknown relationships. When the relationships among the n_1+n_2 sampled

individuals in Figure 4 are completely unknown but are confined to a few candidates such as GG, HS, AN and full-sib (FS), it is still possible to reconstruct the pedigree using unlinked autosomal markers in a likelihood approach (Sieberts et al., 2002; Wang, 2004) if $n_1 > 1$ and $n_2 > 1$. The accuracy of the inferences can be smaller than that shown in Figure 5, but the difference in accuracy should diminish rapidly with an increasing amount of marker information.

Discussion

In marker-assisted parentage analyses, it is common that a mother and a number of her offspring (or fertilized eggs/seeds) are genotyped to infer their paternity (Kichler et al., 1999; Adams et al., 2005; Bretman and Tregenza, 2005; Chapple and Keogh, 2005; Gosselin et al., 2005; Madsen et al., 2005). Some of the offspring sampled from a mother may be full siblings fathered by the same male, and their genotypes can be used jointly to infer the paternity more accurately. Although the full- and HS relationships among the offspring from a mother are usually unknown, they can be identified and utilized to infer their paternity by some recently developed statistical methods (Emery et al., 2001; Sieberts et al., 2002). In such cases, therefore, parentage exclusion probabilities calculated using previous equations assuming a single offspring underestimate the statistical power of parentage analyses and undervalue the amount of information of markers. The equations derived in this study allow more accurate determination of marker information and of the power of parentage analyses. In addition, they can be used to guide experimental designs of parentage analyses in selecting markers and determining the number of offspring to be sampled and genotyped.

Recently, exclusion (Smith et al., 2001; Butler et al., 2004) and likelihood (Thomas and Hill, 2000; Wang, 2004) approaches have been developed to infer sibships in a sample of individuals using their marker genotypes without parental information. To assess the power of and the informativeness of markers in a sibship analysis, Almudevar and Field (1999) derived equations for the probabilities of excluding a number of n-unrelated individuals, a number of $n–1$ full siblings and 1 unrelated (or HS) individual as comprising a full sibship. In the present study, I derived a formula for sibship exclusion probability (S_E) which reduces to very simple forms in some special cases (equations 17–19). I showed that, for a given marker system, S_E increases very rapidly with n, indicating that a group of unrelated individuals is much easily excluded from a full sibship if the group size (n) is large. In other words, an inferred sibship of n individuals becomes increasingly reliable with an increasing value of n, regardless of the methodology (exclusion or likelihood) used in a sibship analysis using a given marker system. The implication for sibship analyses is that the statistical power would be low if most sibships are small in size (say, $n < 4$). In such a case, more informative markers are required to attain sufficient statistical power. On the contrary, when most sibships in a sample of individuals are large, then it is easy to infer the sibships with even a small number of markers.

Traditionally, genealogical relationships or relatedness is inferred between a pair of individuals. Although simple to implement, the pairwise approach suffers from

a number of drawbacks. First, valuable information may be lost in breaking the sampled individuals into pairs and considering each in isolation (Sieberts et al., 2002; Wang, 2004). All individuals in a sample may provide direct and indirect information concerning the relationship of a dyad, especially those closely related to the dyad. In diploid species, for example, sibship exclusion of a group of n individuals is impossible if $n = 2$ but is feasible if $n > 2$ from codominant marker data, as is shown by the present study. Indeed, more accurate relationship inferences are achieved by analysing trios rather than pairs of individuals (Sieberts et al., 2002). This investigation further demonstrates that HS, avuncular and GG relationships can be easily discriminated using unlinked markers when three or more related individuals are analysed jointly. If only a pair of individuals are analysed, however, the three relationships are indistinguishable using unlinked markers, and are only marginally differentiated using linked markers (Epstein et al., 2000). Second, the inferred pairwise relationships are not guaranteed to be self-compatible. Among three individuals, for example, two dyads may be inferred as fullsibs and the other dyad as non-fullsibs from the pairwise methods. The three inferred pairwise relationships are obviously incompatible. In a pairwise parentage analysis, a male and a female may be inferred independently as the father and mother of an offspring, respectively. When the trio are considered jointly, however, the two adults may be incompatible as both parents of the offspring. Third, pairwise approaches infer direct relationships at the lowest level, between a pair of individuals. Such pairwise relationships suffice in some instances in which they are used, for example, to avoid mating between relatives in managing conservation populations (Herbinger et al., 1995). In most cases, however, knowledge of higher order relationships is desirable, which requires all the individuals in a sample to be allocated into various genetic groups (Smith et al., 2001). Further information may be lost in subsequent analyses, such as estimating heritability (Thomas and Hill, 2000), if only pairwise relationships are inferred and used. Although it is possible to first infer pairwise relationships and then cluster them into genetic groups (Blouin et al., 1996; Beyer and May, 2003), such a two-step procedure does not exploit the marker information fully and has to resort to some heuristic rules to resolve the conflicts among some pairwise relationships. This study highlights the great benefits of analysing multiple-related (for example, in inferring parentage or distinguishing GG, HS and AN relationships) or -unrelated (for example, in sibship analyses) individuals to infer their relationships.

I wish to emphasize that parentage (sibship) exclusion probabilities measure adequately the informativeness of markers and the power of parentage (sibship) analyses only when the exclusion approach is adopted in relationship inferences. For other approaches such as likelihood, these probabilities serve the purposes only approximately. In general, a set of markers with a high cumulative exclusion probability and thus a high power in relationship exclusion analyses is also highly informative and gives a high statistical power in likelihood analyses. However, exceptions do exist. For example, biallelic dominant markers such as AFLPs do not allow paternity exclusion in the absence of maternal genotypes (Chakraborty et al., 1974; Gerber et al., 2000). No matter how many

such loci are used, therefore, $P_{E2} \equiv 0$ and the paternity exclusion analysis is powerless. These markers are nevertheless informative and can be used to infer paternity in the likelihood framework. Similarly, biallelic codominant markers such as SNPs are completely uninformative in sibship exclusion analyses ($S_E = 0$) but provide information to differentiate sibship from other relationships by likelihood (Wang, 2006). Some alternative informativeness measurements other than exclusion probabilities have been proposed to measure the information content of markers in inferring genealogical relationships, which apply to all kinds of markers (dominant or codominant, two or more alleles per locus) and relationships and allow for genotyping errors (Wang, 2006).

In the derivation, I followed previous studies in assuming that the markers are in Hardy–Weinberg equilibrium (HWE). It should be noted that this assumption may be violated in real populations, leading to an under- or over-estimation of the exclusion probabilities. A number of conditions are required for a population to reach at and remain in HWE (Crow and Kimura, 1970). Deviation from HWE is resulted when, for example, the marker is under direct or indirect selection, population size is small, mating is not at random with respect to kin (for example, inbreeding avoidance, population subdivision). Whatever the cause of the deviation, its impact on exclusion probability can be formulated using Wright (1965) statistic of F_{IS} denoted by f, following the same approach as adopted in deriving (1), (6), (11) and (16). The formulas become quite complicated, however. For the case of paternity exclusion probability with a known mother, for example, the formula can be derived as

$$
\begin{aligned}
P_{E1} = {}&1 - (f + 2cf_1)f_2a_2 - 2f_1(f_1 - cf_2 - cf_1)a_3 \\
&+ f_1(3(1 - 2c)f_1 + 2cff_2 + 4bf_1f_2)a_4 \\
&+ 2(20b + 11c - 14d)f_1^3a_6 - 4f_1^2(2b + 2c - d \\
&+ (5b - c - d)f_1)(a_5 - a_2a_3) + 8(b + c - d)f_1^3a_2^3 \\
&- 2(6b + 3c - 4d)f_1^3a_3^2 - 4(9b + 6c - 7d)f_1^3a_2a_4 \\
&- 2f_1(f_1 + cf + (2b - c)f_1f_2)a_2^2
\end{aligned}
$$

where $f_1 = 1 - f$ and $f_2 = 2 - f$, and a, b, c and d are as defined in (1). When the marker is in HWE so that $f = 0$, the above formula reduces to (1) as expected. The impact of f on P_{E1}
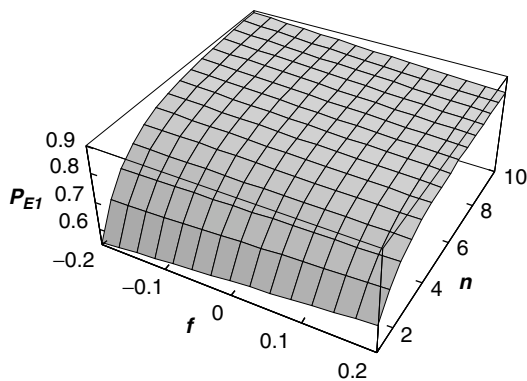

Figure 6 The effect of deviation from Hardy–Weinberg equilibrium (f) on the paternity exclusion probability (P_{E1}) when a known mother and n full-sib offspring are used in a paternity analysis. A single locus with five equal frequency codominant alleles is assumed in the calculation.

is shown in Figure 6 for a locus with five codominant alleles of an equal frequency. It can be seen that the magnitude of effect on P_{E1} of the deviation from HWE is relatively small, and that the direction of effect on P_{E1} depends on n. When n is small, inbreeding (positive f) leads to an increase in P_{E1}, while when n is large, inbreeding results in a decrease in P_{E1}. As an empirical example, consider the marker with 14 codominant alleles with frequencies listed in the first row of Table 2. When $n = 1$, the values of P_{E1} are 0.7759, 0.7873, 0.7985 with $f = -0.1, 0, 0.1$, respectively. When $n = 4$, the values of P_{E1} become 0.9430, 0.9336, 0.9251 with $f = -0.1, 0, 0.1$, respectively.

The assumption of linkage equilibrium (LE) is required to calculate multi-locus exclusion probabilities simply from single-locus values. Unlike HWE, it is difficult to relax this assumption in deriving the exclusion probabilities. However, like HWE, slight deviations from LE should have a small effect on exclusion probabilities. To investigate quantitatively the impact of deviation from LE, a further simulation study should be conducted.

References

Adams EM, Jones AG, Arnold SJ (2005). Multiplepaternityin a natural population of a salamander with long-term sperm storage. *Mol Ecol* **14**: 1803–1810.

Almudevar A, Field C (1999). Estimation of single-generation sibling relationships based on DNA markers. *J Agric Biol Env Stat* **4**: 136–165.

Avise JC, Jones AG, Walker D, DeWoody JA (2002). Genetic mating systems and reproductive natural histories of fishes: lessons for ecology and evolution. *Annu Rev Genet* **36**: 19–45.

Beyer J, May B (2003). A graph-theoretic approach to the partition of individuals into full-sib families. *Mol Ecol* **12**: 2243–2250.

Blouin MS, Parsons M, Lacaille V, Lotz S (1996). Use of microsatellite loci to classify individuals by relatedness. *Mol Ecol* **5**: 393–401.

Bretman A, Tregenza T (2005). Measuring polyandry in wild populations: a case study using promiscuous crickets. *Mol Ecol* **14**: 2169–2179.

Butler K, Field C, Herbinger CM, Smith BR (2004). Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Mol Ecol* **13**: 1589–1600.

Chakraborty R, Meagher TR, Smouse PE (1988). Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. *Genetics* **118**: 527–536.

Chakraborty R, Shaw M, Schull WJ (1974). Exclusion of probability: the current state of the art. *Am J Hum Genet* **26**: 477–488.

Chapple DG, Keogh JS (2005). Complex mating system and dispersal patterns in a social lizard, *Egernia whitii*. *Mol Ecol* **14**: 1215–1227.

Coltman DW, Bancroft DR, Robertson A, Smith JA, Clutton-Brock TH, Pemberton JM (1999). Male reproductive success in a promiscuous mammal: behavioural estimates compared with genetic paternity. *Mol Ecol* **8**: 1199–1209.

Crow JF, Kimura M (1970). *An Introduction to Population Genetics Theory*. Harper and Row: New York.

Dodds KG, Tate ML, McEwan JC, Crawford AM (1996). Exclusion probabilities for pedigree testing farm animals. *Theor Appl Genet* **92**: 966–975.

Double MC, Cockburn A, Barry SC, Smouse PE (1997). Exclusion probabilities for single-locus paternity analysis when related males compete for matings. *Mol Ecol* **6**: 1155–1166.

Emery AM, Wilson IJ, Craig S, Boyle PR, Noble LR (2001). Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Mol Ecol* **10**: 1265–1278.

Epstein MP, Duren WL, Boehnke M (2000). Improved inference of relationship for pairs of individuals. *Am J Hum Genet* **67**: 1219–1231.

Fung WK, Chung YK, Wong DM (2002). Power of exclusion revisited: probability of excluding relatives of the true father from paternity. *Int J Legal Med* **116**: 64–67.

Garant D, Dodson JJ, Bernatchez L (2001). A genetic evaluation of mating system and determinants of individual reproductive success in Atlantic salmon (*Salmo salar* L.). *J Hered* **92**: 137–145.

Gerber S, Mariette S, Streiff R, Bodenes C, Kremer A (2000). Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. *Mol Ecol* **9**: 1037–1048.

Gosselin T, Sainte-Marie B, Bernatchez L (2005). Geographic variation of multiple paternity in the American lobster, *Homarus americanus*. *Mol Ecol* **14**: 1517–1525.

Herbinger CM, Doyle RW, Pitman ER, Paquet D, Mesa KA *et al.* (1995). DNA fingerprint based analysis of paternal and maternal effects on offspring growth and survival in communally reared rainbow trout. *Aquaculture* **137**: 245–256.

Hu YQ, Fung WK, Hu YQ (2005). Power of excluding an elder brother of a child from paternity. *Forensic Sci Int* **152**: 321–322.

Hughes CR (1998). Integrating molecular techniques with field methods in studies of social behavior: a revolution results. *Ecology* **79**: 383–399.

Jamieson A (1965). The genetics of transferrin in cattle. *Heredity* **20**: 419–441.

Jamieson A, Taylor SCS (1997). Comparisons of three probability formulae for parentage exclusion. *Anim Genet* **28**: 397–400.

Jones AG (2001). Gerud 1.0: a computer program for the reconstruction of parental genotypes from progeny arrays using multilocus DNA data. *Mol Ecol Notes* **1**: 215–218.

Jones AG, Ardren WR (2003). Methods of parentage analysis in natural populations. *Mol Ecol* **12**: 2511–2523.

Kichler K, Holder MT, Davis SK, Marquez R, Owens DW (1999). Detection of multiple paternity in the Kemp's ridley sea turtle with limited sampling. *Mol Ecol* **8**: 819–830.

Madsen T, Ujvari B, Olsson M, Shine R (2005). Paternal alleles enhance female reproductive success in tropical pythons. *Mol Ecol* **14**: 1783–1787.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Mol Ecol* **7**: 639–655.

McPeek MS, Sun L (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet* **66**: 1076–1094.

Ohno Y, Sebetan IM, Akaishi S (1982). A simple method for calculating the probability of excluding paternity with any number of codominant alleles. *Forensic Sci Int* **19**: 93–98.

Painter I (1997). Sibship reconstruction without parental information. *J Agric Biol Env Stat* **2**: 212–229.

Robledo-Arnuncio JJ, Gil L (2005). Patterns of pollen dispersal in a small population of *Pinus sylvestris* L. revealed by total-exclusion paternity analysis. *Heredity* **94**: 13–22.

Salmon DB, Brocteur J (1978). Probability of paternity exclusion when relatives are involved. *Am J Hum Genet* **30**: 65–75.

Selvin S (1980). Probability of non-paternity determined by multiple allele codominant systems. *Am J Hum Genet* **32**: 276–278.

Sieberts SK, Wijsman EM, Thompson EA (2002). Relationship inference from trios of individuals, in the presence of typing error. *Am J Hum Genet* **70**: 170–180.

Smith BR, Herbinger CM, Merry HR (2001). Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics* **158**: 1329–1338.

Thomas SC, Hill WG (2000). Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* **155**: 1961–1972.

Thomas SC, Hill WG (2002). Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genet Res* **79**: 227–234.

Thompson EA, Meagher TR (1987). Parental and sib likelihoods in genealogy reconstruction. *Biometrics* **43**: 585–600.

Villanueva B, Verspoor E, Visscher PM (2002). Parental assignment in fish using microsatellite genetic markers with finite numbers of parents and offspring. *Anim Genet* **33**: 33–41.

Wang JL (2004). Sibship reconstruction from genetic data with typing errors. *Genetics* **166**: 1963–1979.

Wang JL (2006). Informativeness of genetic markers for pairwise relationship and relatedness inference. *Theor Popul Biol* **70**: 300–321.

Weir BS (1996). *Genetic Data Analysis II*. Sinauer Associates: Sunderland, MA.

Wiener AS, Lederer M, Polayes SH (1930). Studies in isohemagglutination IV: on the chances of proving non-paternity: with special reference to blood groups. *J Immunol* **19**: 259–282.

Wright S (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**: 395–420.

Appendix

Deriving sibship exclusion probability

To derive S_E, I first consider the probability (T_E) that n unrelated individuals cannot be excluded from a sibship, and $1-T_E$ gives S_E. Denote the genotypes of n-unrelated individuals at a k-allele locus as $\mathbf{G} = \{G_1, G_2, \ldots, G_n\}$. $\mathbf{G}$ does not allow sibship exclusion in the following cases.

1. $\mathbf{G}$ contains one allele

In this case all of the n individuals display the same homozygous genotype, with a probability $T_{E,1} = \sum_u (p_u^2)^n$

2. $\mathbf{G}$ contains two alleles

The probability is

$$T_{E,2} = \frac{1}{2} \sum_u \sum_{v \neq u} \sum_{i=1}^{2n-1} \frac{(2n)!}{i!(2n-i)!} p_u^i p_v^{2n-i}$$

3. $\mathbf{G}$ contains three alleles observed in two or three kinds of heterozygotes

In this case, $\mathbf{G}$ consists of genotypes $\{A_u A_v, A_u A_w\}$ or $\{A_u A_v, A_u A_w, A_v A_w\}$, where $u \neq v \neq w = 1, \ldots, k$. The probability is

$$T_{E,3} = \frac{1}{2} \sum_u \sum_{v \neq u} \sum_{w \neq u,v} \sum_{i=1}^{n-1} \frac{n!}{i!(n-i)!} \times (2p_u p_v)^i (2p_u p_w)^{n-i}$$

$$+ \frac{1}{6} \sum_u \sum_{v \neq u} \sum_{w \neq u,v} \sum_{i=1}^{n-2} \sum_{j=1}^{n-i-1} \frac{n!}{i!j!(n-i-j)!}$$

$$\times (2p_u p_v)^i (2p_u p_w)^j (2p_v p_w)^{n-i-j}$$

4. $\mathbf{G}$ contains three alleles observed in one kind of homozygote and one or more kinds of heterozygotes

In this case, $\mathbf{G}$ consists of genotypes $\{A_u A_u, A_v A_w\}$, or $\{A_u A_u, A_v A_w, A_u A_v\}$, or $\{A_u A_u, A_v A_w, A_u A_w\}$, or $\{A_u A_u, A_u A_v, A_u A_w\}$, or $\{A_u A_u, A_v A_w, A_u A_v, A_u A_w\}$ where $u \neq v \neq w = 1, \ldots, k$. The probability of the case is

$$T_{E,4} = \frac{1}{2} \sum_u \sum_{v \neq u} \sum_{w \neq u,v} \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} \sum_{s=0}^{n-i-j} \frac{n!}{i!j!s!(n-i-j-s)!}$$
$$\times (p_u^2)^i (2p_v p_w)^j (2p_u p_v)^s (2p_u p_w)^{n-i-j-s}$$
$$+ \frac{1}{2} \sum_u \sum_{v \neq u} \sum_{w \neq u,v} \sum_{i=1}^{n-2} \sum_{j=1}^{n-i-1} \frac{n!}{i!j!(n-i-j)!}$$
$$\times (p_u^2)^i (2p_u p_v)^j (2p_u p_w)^{n-i-j}$$

5. **G** contains four alleles observed in two kinds of heterozygotes

The genotypes in **G** are {$A_u A_v$, $A_w A_x$}, where $u \neq v \neq w \neq x = 1, \ldots, k$. The probability of the case is

$$T_{E,5} = \frac{1}{8} \sum_u \sum_{v \neq u} \sum_{w \neq u,v} \sum_{x \neq u,v,w} \sum_{i=1}^{n-1} \frac{n!}{i!(n-i)!}$$
$$\times (2p_u p_v)^i (2p_w p_x)^{n-i}$$

6. **G** contains four alleles observed in three kinds of heterozygotes that do not share an allele among all of them

A possible genotype combination in **G** is {$A_u A_v$, $A_u A_w$, $A_v A_x$}, where $u \neq v \neq w \neq x = 1, \ldots, k$. The probability of the case is

$$T_{E,6} = \frac{1}{2} \sum_u \sum_{v \neq u} \sum_{w \neq u,v} \sum_{x \neq u,v,w} \sum_{i=1}^{n-2} \sum_{j=1}^{n-i-1} \frac{n!}{i!j!(n-i-j)!}$$
$$\times (2p_u p_v)^i (2p_u p_w)^j (2p_w p_x)^{n-i-j}$$

7. **G** contains four alleles observed in four kinds of heterozygotes that do not share an allele among any three of them

A possible genotype combination in **G** is {$A_u A_v$, $A_u A_w$, $A_v A_x$, $A_w A_x$}, where $u \neq v \neq w \neq x = 1, \ldots, k$. The probability of the case is

$$T_{E,7} = \frac{1}{8} \sum_u \sum_{v \neq u} \sum_{w \neq u,v} \sum_{x \neq u,v,w} \sum_{i=1}^{n-3} \sum_{j=1}^{n-i-2}$$
$$\times \sum_{s=1}^{n-i-j-1} \frac{n!}{i!j!s!(n-i-j-s)!}$$
$$\times (2p_u p_v)^i (2p_u p_w)^j$$
$$\times (2p_v p_x)^s (2p_w p_x)^{n-i-j-s}$$

The total non-exclusion probability, T_E, is obtained by summing T_{Ei} for $i = 1, 2, \ldots, 7$. After some tedious algebra, $S_E = 1 - T_E$ becomes (16) in text.